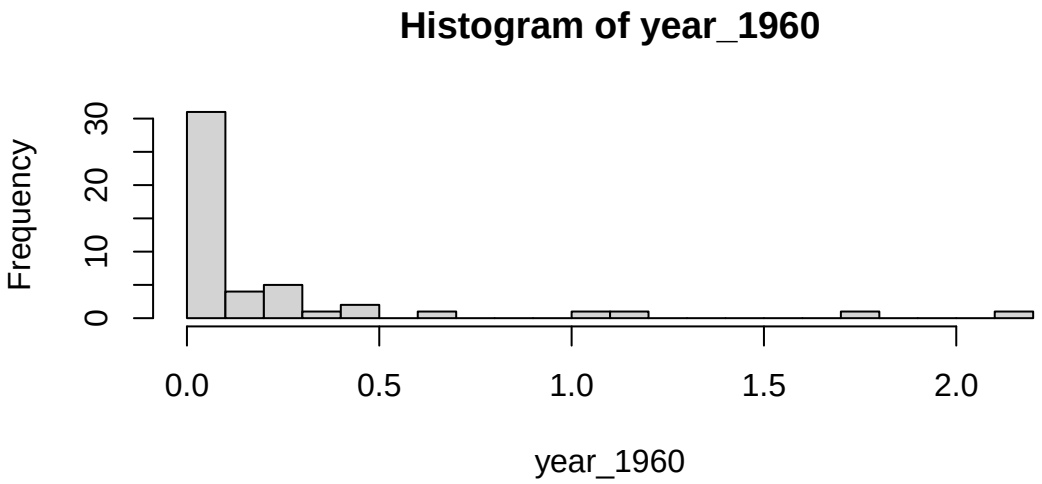
Final Project

Shuning Zhu

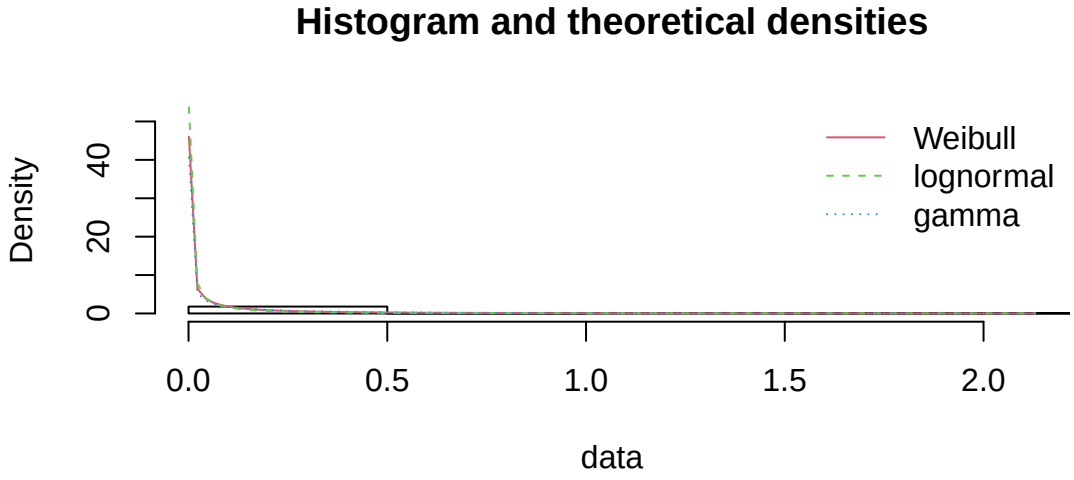
5/2/2022

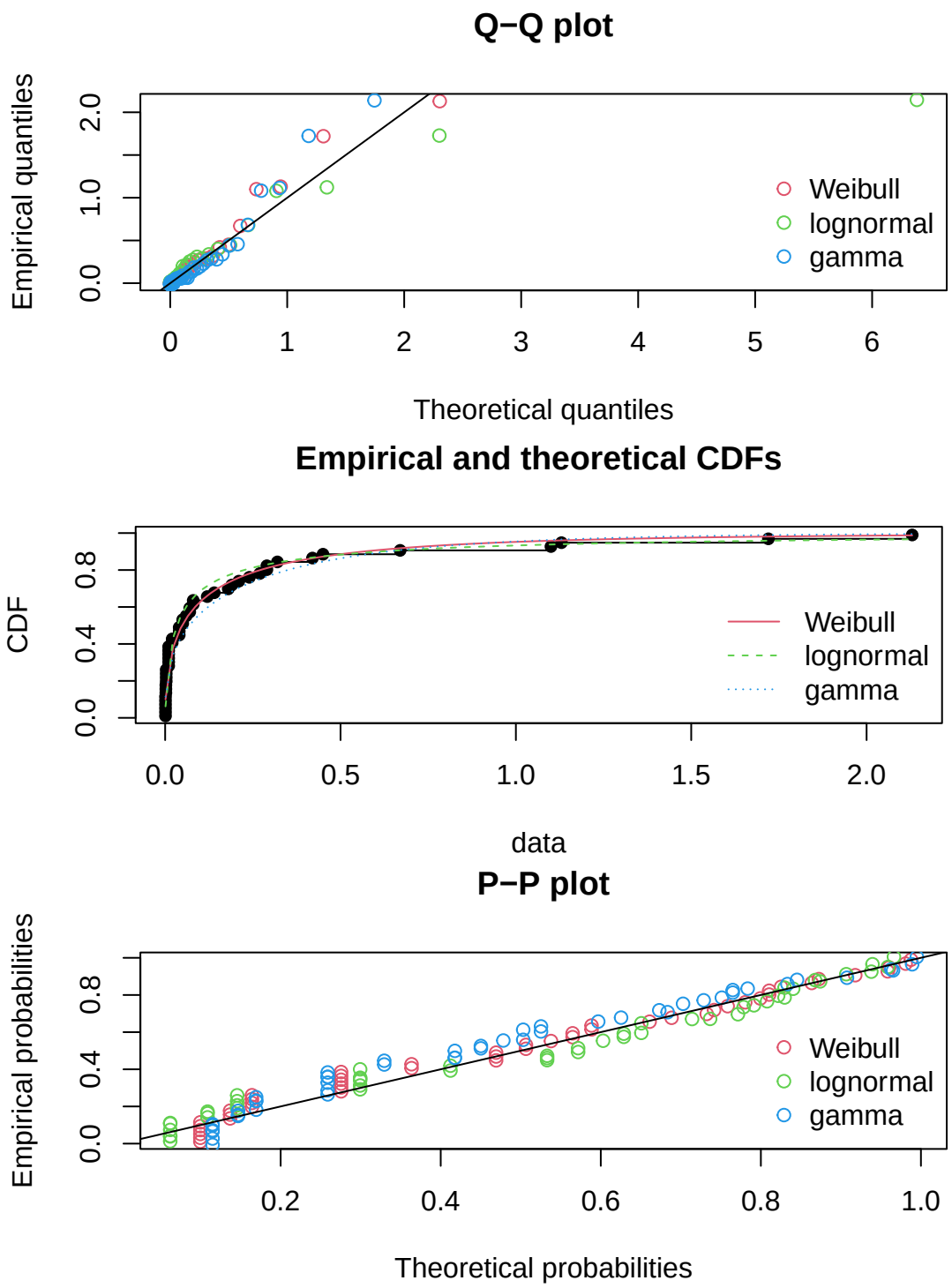
Illinois_rain distribution

Fit distribution for each year



I assume that rainfall data of all the 5 years share one common distribution but with different parameter values. After looking into the histograme of the rainfall data of year 1960 I notice that it's obviously left-bounded which means all the records are more than 0. And it's also right-skewed distributed with a tail on the right side. So I consider to use some specific distributions like weibull, gamma, lognorm distribution to fit the data. The choose one best distribution to describe the rainfall data.





Plots above are four Goodness-of-fit plots for various distributions fitted to continuous data (Weibull, gamma and lognormal distributions fitted to rainfall dataset) as provided by functions denscomp, qqcomp, cdfcomp and ppcomp. To start with, I look into the Histogram vs Theoretical densities. However I can not find significant difference among these three distributions. So then I look into the p-p plot to see the goodness of these three distributions fit to the center of the data set and it shows that the Weibull distribution has a relatively outstanding performance on that. Also I check the q-q plot and it shows that none of these three distributions fit the right tail of the data very well although the left tail is well fitted by all of them. I find it hard to tell which distribution fits the data better than others directly from the visualizations so then I am

going to look into and compare some Goodness-of-fit statistics of these distributions.

```
## Goodness-of-fit statistics
##                                 Weibull lognormal     gamma
## Kolmogorov-Smirnov statistic 0.12026560 0.1370492 0.1251773
## Cramer-von Mises statistic   0.08277497 0.1816660 0.1131148
## Anderson-Darling statistic   0.64480151 1.1608528 0.7801751
##
## Goodness-of-fit criteria
##                                  Weibull  lognormal     gamma
## Akaike's Information Criterion -105.4157 -100.84151 -107.8918
## Bayesian Information Criterion -101.6733  -97.09911 -104.1494
```

It can be found from the result above that weibull distribution has an outstanding performance in terms of Kolmogorov-Smirnov statistic, Cramer-von Mises statistic and Anderson-Darling statistic. Gamma distribution has an outstanding performance in terms of AIC and BIC. Personally I choose Weibull distribution to do following steps.

Parameter estimation using MLE

```
## Fitting of the distribution ' weibull ' by maximum likelihood
## Parameters :
##        estimate Std. Error
## shape 0.4872196 0.05403717
## scale 0.1021156 0.03201623
## Loglikelihood:  54.70784   AIC:  -105.4157   BIC:  -101.6733
## Correlation matrix:
##          shape    scale
## shape 1.000000 0.327801
## scale 0.327801 1.000000
```

```
## Fitting of the distribution ' weibull ' by maximum likelihood
## Parameters :
##        estimate Std. Error
## shape 0.6931890 0.07953545
## scale 0.2171053 0.04764433
## Loglikelihood:  20.06308   AIC:  -36.12616   BIC:  -32.38376
## Correlation matrix:
##           shape     scale
## shape 1.0000000 0.3161267
## scale 0.3161267 1.0000000
```

```
## Fitting of the distribution ' weibull ' by maximum likelihood
## Parameters :
##        estimate Std. Error
## shape 0.5449274 0.05623383
## scale 0.1049731 0.02720642
## Loglikelihood:  62.21192   AIC:  -120.4238   BIC:  -116.3731
## Correlation matrix:
##           shape     scale
## shape 1.0000000 0.3245771
## scale 0.3245771 1.0000000
```

```
## Fitting of the distribution ' weibull ' by maximum likelihood
## Parameters :
##        estimate Std. Error
```

```
## shape 0.6475664 0.08363603
## scale 0.1919355 0.05143966
## Loglikelihood:  19.46213   AIC:  -34.92426   BIC:  -31.70243
## Correlation matrix:
##           shape     scale
## shape 1.0000000 0.3203635
## scale 0.3203635 1.0000000
```

```
## Fitting of the distribution ' weibull ' by maximum likelihood
## Parameters :
##        estimate Std. Error
## shape 0.5701875 0.07274632
## scale 0.1175969 0.03537006
## Loglikelihood:  38.11581   AIC:  -72.23163   BIC:  -68.95646
## Correlation matrix:
##           shape     scale
## shape 1.0000000 0.3250065
## scale 0.3250065 1.0000000
```

The tables above show the estimations of parameters of 5 years' rainfall data using MLE.

Identify wet and dry years

My next step is to identify identify wet years and dry years using this distribution,

```
##   Years        sd total_Rainfall mean_rainfall Storm_num   type
## 1  1960 0.4391915         10.574     0.2458649        48 normal
## 2  1961 0.3706052         13.197     0.2539730        48    wet
## 3  1962 0.3493277         10.346     0.1637297        56 normal
## 4  1963 0.3726197          9.710     0.2624324        37 normal
## 5  1964 0.2699068          7.110     0.1908108        38    dry
```

I define the wet year as in which the total rainfall is an upper outlier of these 5 years' total rainfall. I define the dry year as in which the total rainfall is an lower outlier of these 5 years' total rainfall. The years in the middle are normal years.

More storms nor more rainfall brought by individual storm does not separately correspond to more rainfall.For instance, 1962, the year with most storms, is a normal year. Further more, fewer storms happened in 1963 than 1964, nevertheless the later year is a dry year. But the product of them is the total rainfall which directly determines if a specific year is wet or dry or normal.

Extent

The article by Floyd Huff discussed that the individual effects of mean rainfall, storm duration,and other storm factors were small and erratic in behavior when the foregoing analytical technique was used. As a result, we don't have enough confidence to claim that the storm has no relationship with rainfall due to the small data set. What we can extent in next step is collecting enough data to make a more solid conclusion.

4.25

```
## [1] 0.4166667
```

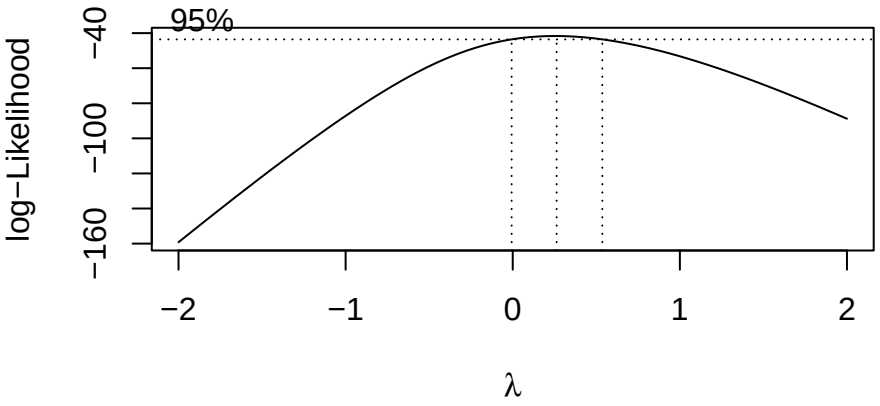
```
## [1] 0.40625
```

```
## [1] 0.4545455
```

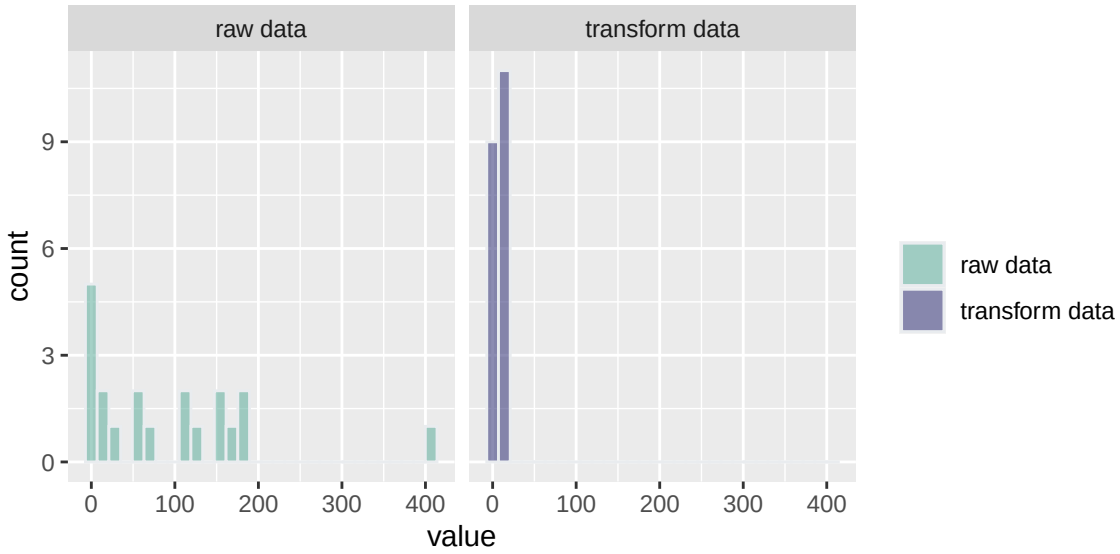
```
## [1] 0.4516129
```

We can respectively obtain expectations and medians, and find that expectations are approximately equal to the medians.

4.39



We can make comparison between the raw data and the transform data by the histogram plot.



##4.27

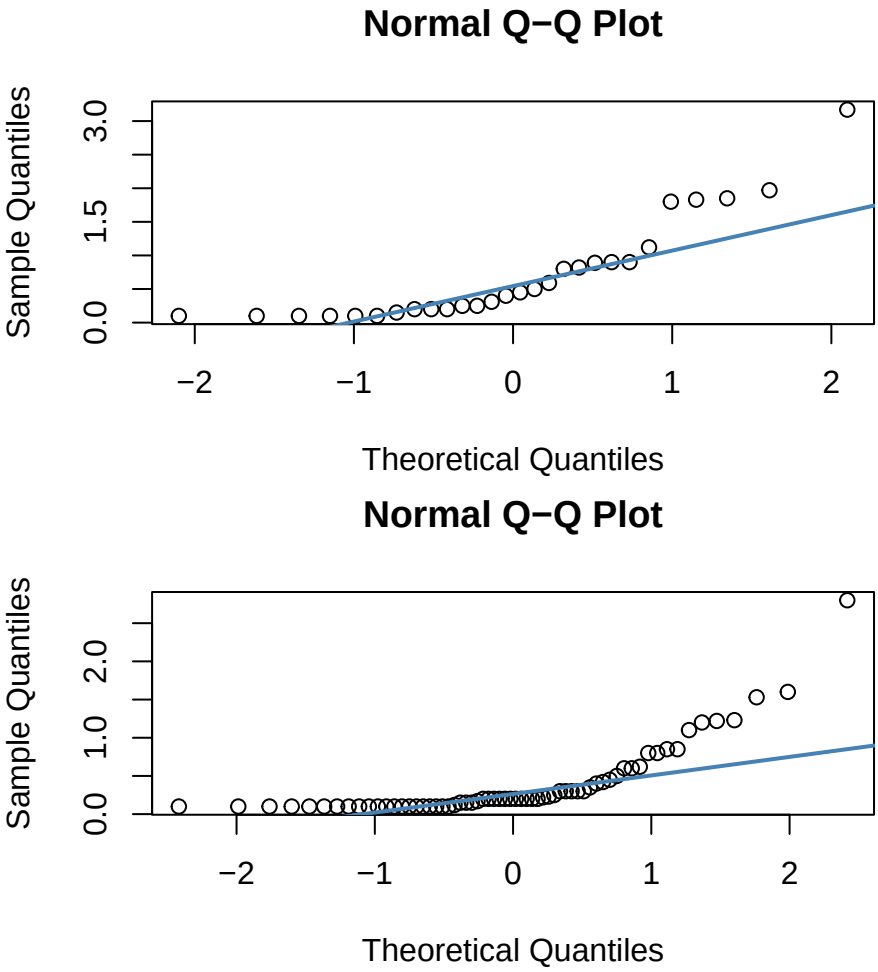
(a)

```
##    vars  n mean   sd median trimmed  mad min  max range skew kurtosis   se
## X1    1 28 0.72 0.77   0.43    0.62 0.48 0.1 3.17  3.07 1.47     1.63 0.14
```

```
##    vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 64 0.39 0.48    0.2    0.29 0.15 0.1 2.8   2.7 2.64     8.41 0.06
```

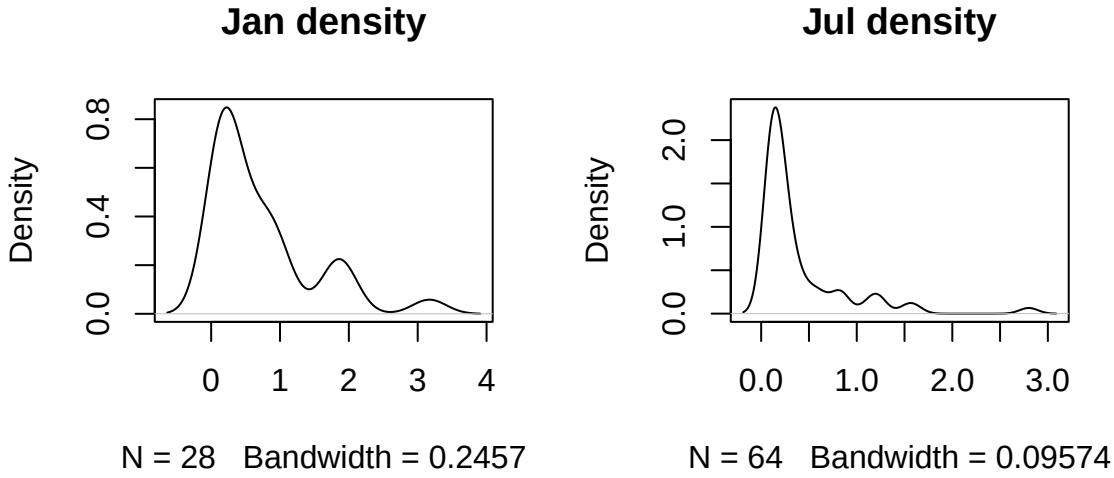
Based on the summary statistics from two data sets, we can conclude that data set Jan contains higher mean, median, max and range values, whereas data set Jul contains more variables and higher skew value.

(b)



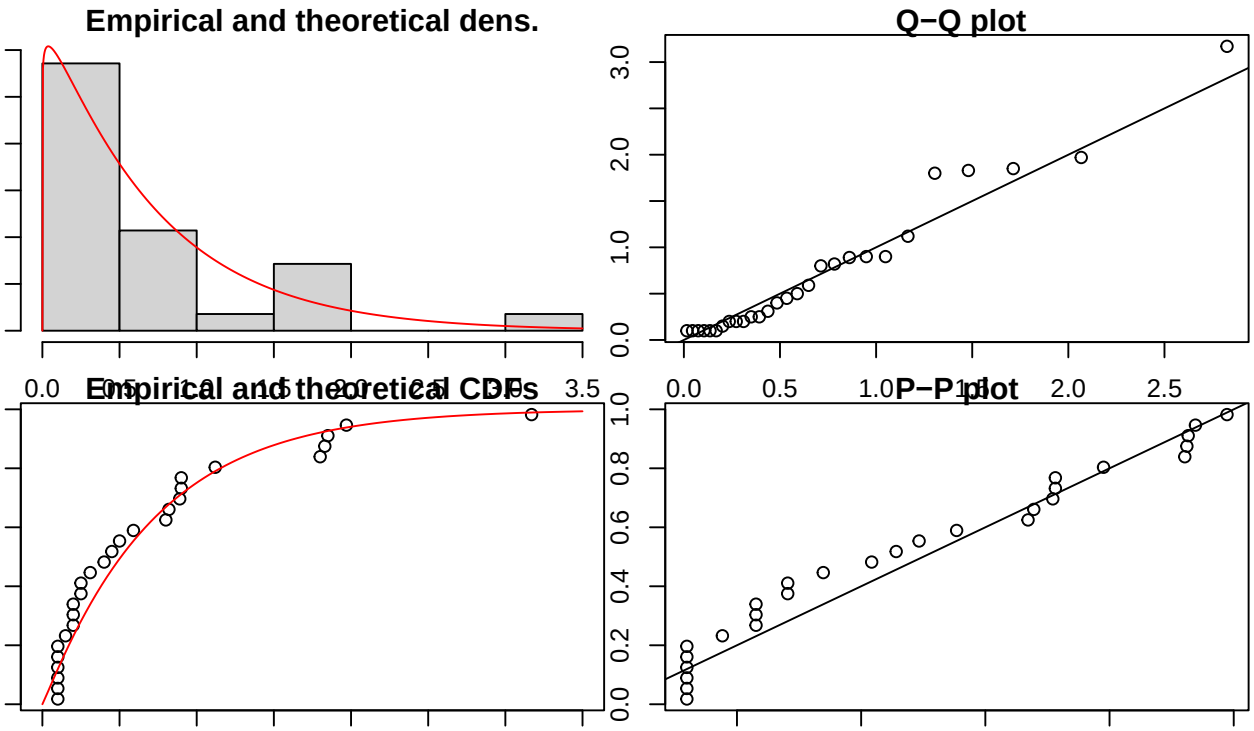
The qqplots have light tails, as the result, we think the normal distribution is unreasonable for this problem.

We generate density plots to prove the conclusion. The distributions are closer to the gamma distribution rather than normal distribution.



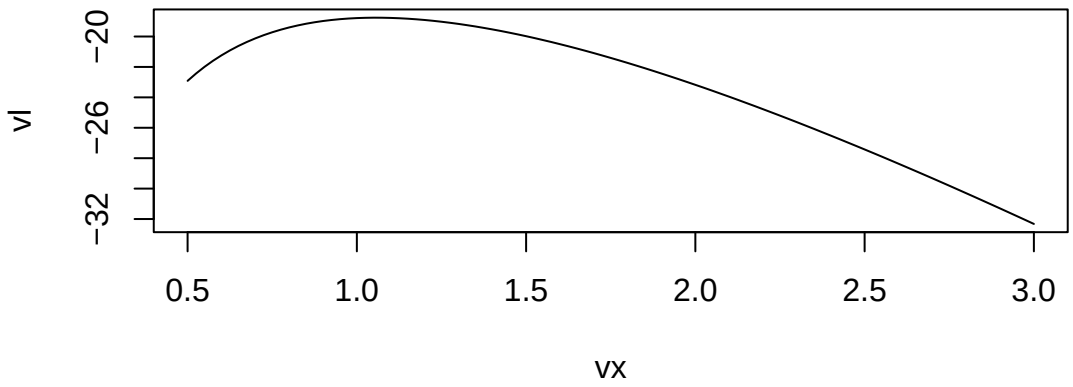
(c)

```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## shape 1.056222  0.2497495
## rate  1.467650  0.4396202
## Loglikelihood:  -18.7616   AIC:  41.5232   BIC:  44.18761
## Correlation matrix:
##           shape      rate
## shape 1.0000000 0.7893943
## rate  0.7893943 1.0000000
```



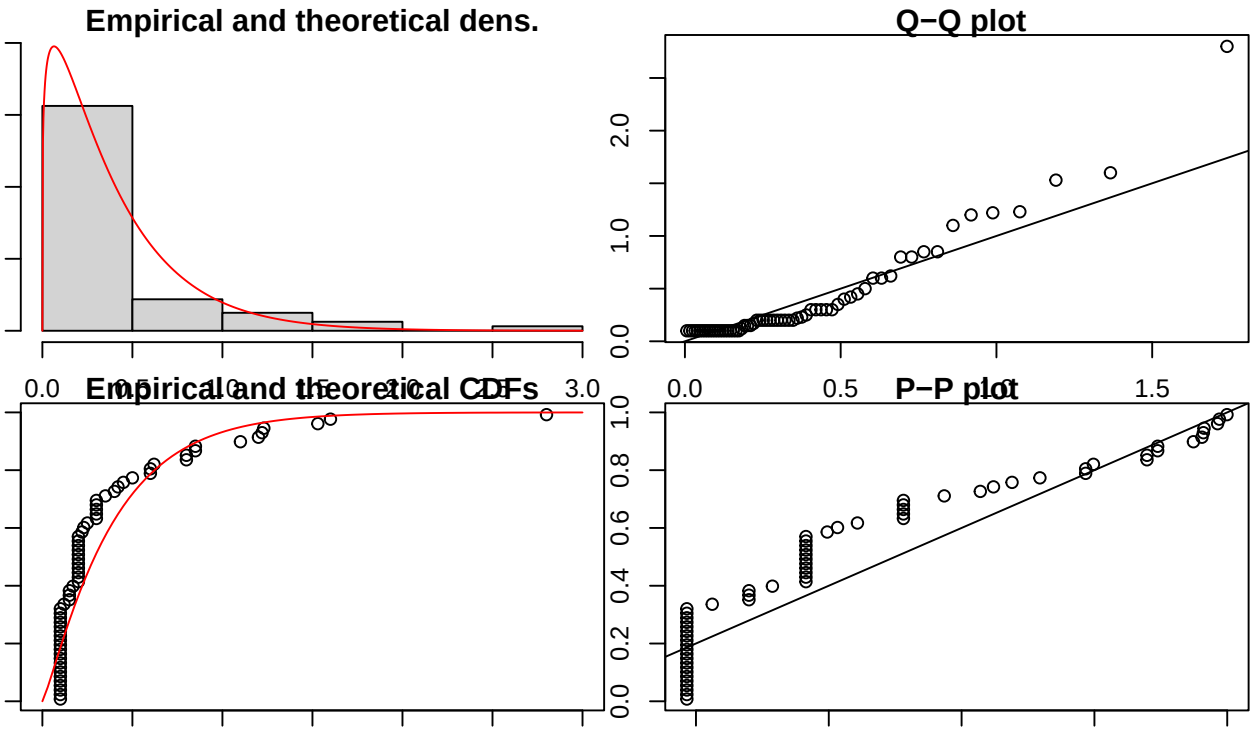

```
## [1] 0.5667132 1.5457314
```

```
## $par
## [1] 1.056378 1.467814
##
## $value
## [1] 18.7616
##
## $counts
## function gradient
##       55       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

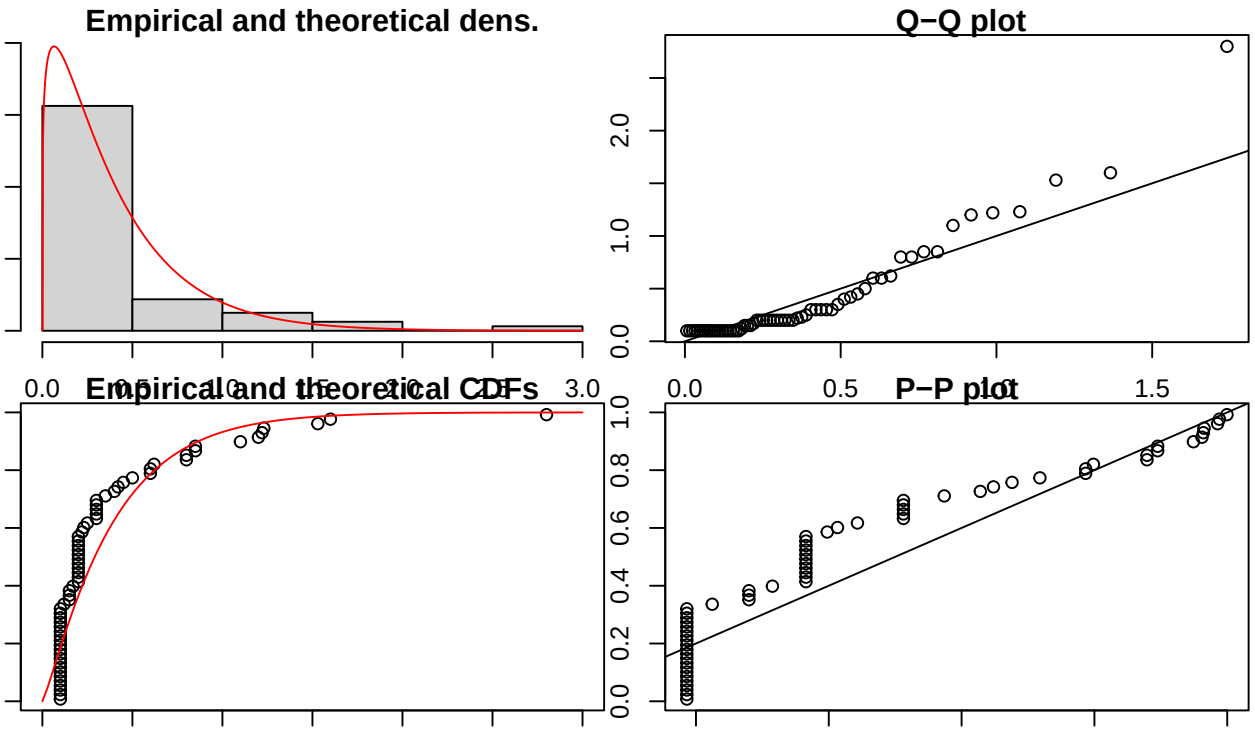



```
## $par
## [1] 1.05625
##
## $value
## [1] 18.7616
##
## $counts
## function gradient
##       20       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## shape 1.196419  0.1891196
## rate  3.043403  0.5936302
## Loglikelihood:  -3.634886   AIC:  11.26977   BIC:  15.58754
## Correlation matrix:
##           shape      rate
## shape 1.0000000 0.8103948
## rate  0.8103948 1.0000000
```

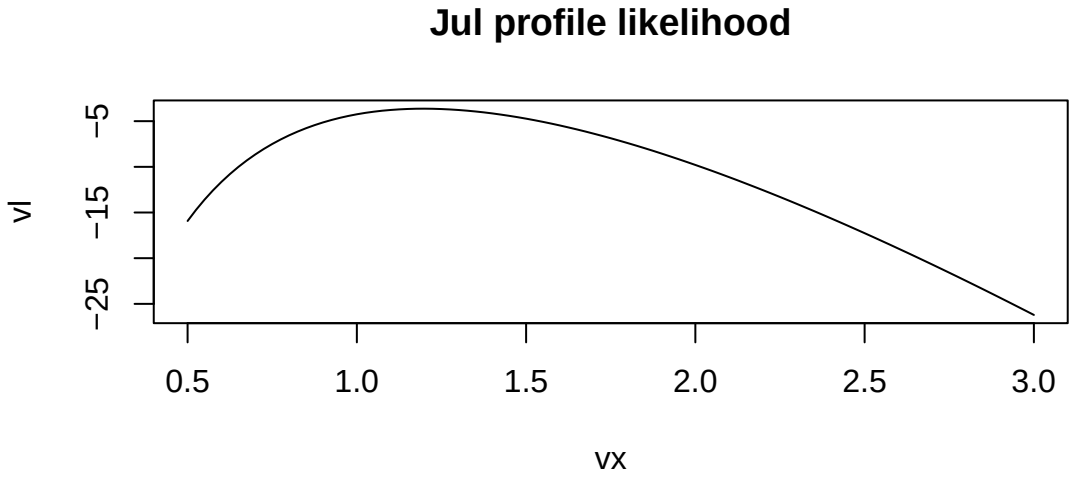



```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## shape 1.196419  0.1891196
## rate  3.043403  0.5936302
## Loglikelihood:  -3.634886   AIC:  11.26977   BIC:  15.58754
## Correlation matrix:
##           shape      rate
## shape 1.0000000 0.8103948
## rate  0.8103948 1.0000000
```




```
## [1] 0.8257444 1.5670931
```

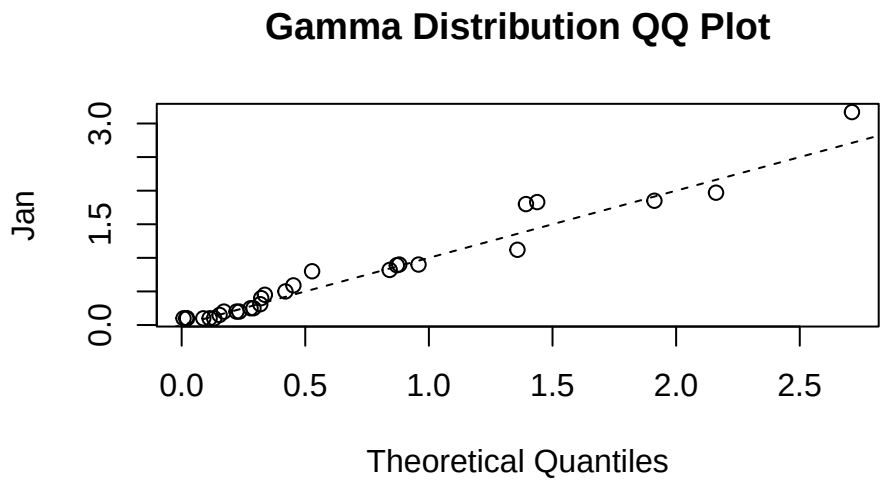
```
## $par
## [1] 1.196268 3.042774
##
## $value
## [1] 3.634887
##
## $counts
## function gradient
##       71       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

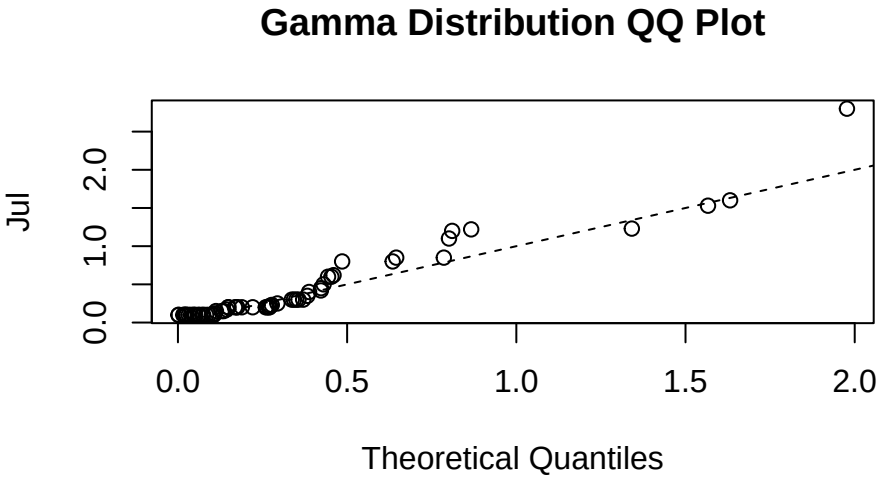



```
## $par
## [1] 1.196289
##
## $value
## [1] 3.634887
##
## $counts
## function gradient
##       22       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

Compare the parameters, Jul data set has higher maximum likelihood estimator, and it fits better.

(d)





It seems that Jul data set fits better in gamma distribution.